Parametric and semi-parametric modelling of vacation expenditures

Melenberg, B.; van Soest, A.H.O.

Published in:
Journal of Applied Econometrics

Publication date:
1996

JOURNAL OF APPLIED ECONOMETRICS, VOL. 11, 59–76 (1996)

PARAMETRIC AND SEMI-PARAMETRIC MODELLING OF VACATION EXPENDITURES

BERTRAND MELENBERG AND ARTHUR VAN SOEST
Tilburg University, Department of Econometrics, PO Box 90153, 5000 LE Tilburg, The Netherlands

SUMMARY

We analyse several limited dependent variable models explaining the budget share that Dutch families spend on vacations. To take account of the substantial number of zero shares, two types of models are used. The first is the single-equation censored regression model. We estimate and test several parametric and semi-parametric extensions of the Tobit model. Second, we consider two-equation models, in which the participation decision and the decision on the amount to spend are treated separately. The first decision is modelled as a binary choice model; the second as a conditional regression. We estimate and test parametric and semi-parametric specifications.

1. INTRODUCTION

In this paper we analyse annual vacation expenditures of Dutch families. Using cross-section data, we study the impact of family income, household composition, and other family characteristics, on the share of total family expenditures spent on vacations.

During the last decades, the tourist market has grown substantially in many countries (cf. OECD, 1989). In the Netherlands, the fraction of families with at least one vacation per year gradually increased from 40·8% in 1966 to 62·6% in 1981, and then slightly fell to 60·3% in 1985 (OECD, 1989). From 1970 to 1981, vacation expenditures increased from 3·5% to 4·0% of total consumer expenditures. This suggests that it is useful to analyse the factors determining household vacation expenditures. Studying total vacation expenditures can be seen as a first step towards a more disaggregate analysis, accounting for the heterogeneous nature of vacation goods. For example, Eymann and Ronning (1992) used nested models, with the vacation participation decision at the first level, and the choice between a number of geographical destinations in a second step. A micro model for vacation expenditures can also be embedded in a complete model for the Dutch tourist market, for example used by the Dutch government (SEO, 1991).

A typical characteristic of expenditures on vacations compared to other expenditure categories like food, clothing, etc. is that many families do not spend anything on vacations. In our sample, this is the case for 37% of all observations. To account for this, two types of limited dependent variable models are adequate: the one-equation censored regression model, of which the familiar Tobit model is the standard special case, and a two-equations model, with a binary choice part explaining the participation decision, and a conditional regression equation explaining the expenditure level if this is positive.[1]

[1] 'To go on holiday', 'to participate', and 'positive vacations budget share' are used as synonyms. A family participates if at least one of its members goes on holiday.

CCC 0883–7252/96/010059–18

Received December 1991
Revised March 1995

Standard estimators for these types of models are based on maximum likelihood (ML) or two-step procedures. Their consistency generally requires a complete and correct specification of a parametric family of the error distribution. This first implies the need of thorough specification testing. Many tests are available in a rather general framework. See, for example, the studies in Blundell (1987). If the model appears to be misspecified, model assumptions must be relaxed, and estimators are needed which remain consistent under more general assumptions. Various semi-parametric models and estimators have recently been developed for this goal (cf. Robinson, 1988, for a survey), but applications are still sparse.

In this paper we focus on thorough model selection and semi-parametric estimation. Following the strategy of Horowitz (1993), we estimate, test, and compare parametric and semi-parametric models. Our methodological aim is to establish the practical value of some recently developed estimation techniques, and to compare them with more traditional methods. This is of interest for many similar models in which a mixed discrete continuous decision is analysed. See, for example, the discussion on zero observations in household consumer expenditures in Blundell (1990).

In Section 2 we consider the single-equation censored regression model. The starting point is the Tobit model, characterized by homoscedasticity and normality of the errors. Pseudo-ML estimates obtained by maximizing the Tobit likelihood are generally inconsistent if homoscedasticity or normality is violated. Our test results imply that both assumptions are rejected. Moreover, parametric extensions allowing for heteroscedasticity and non-normality are also rejected. We then consider two semi-parametric estimators which allow for both heteroscedasticity and asymmetry: the censored least absolute deviations estimator (Powell, 1984) and an efficient two-step estimator given by Newey and Powell (1990).

Apart from many observations with zero expenditures on vacations, our data lack observations with small positive vacation expenditures. This is not captured in the usual (zero threshold) censored regression model. In Van Soest and Kooreman (1987) a model with unobserved random thresholds was used to take account of this. For such a model, semi-parametric estimators are not yet available. We consider an alternative model with similar flexibility. In Section 3, vacation expenditures are modelled in two steps: a binary choice equation to model participation, and, for the non-zero observations, a regression equation to explain the expenditure level. We consider semi-parametric specifications of the binary choice model, but cannot reject the Probit model. For the conditional regression equation, normality and homoscedasticity are rejected. We compare OLS to the estimator of Robinson (1987), which achieves the efficiency bound for the weak identifying assumption of zero conditional error means. In Section 4, we briefly compare the models of Sections 2 and 3.

2. THE CENSORED REGRESSION MODEL

In this section we use a censored regression model to explain family expenditures on vacations as a function of total expenditures and other family characteristics:

$$y_i^* = X_i'\alpha + \varepsilon_i, \qquad y_i = \max(y_i^*, 0) \tag{1}$$

Here y_i is the budget share family i spends on vacations ($i = 1, \ldots, N$), i.e. annual vacation expenditures as a percentage of total annual expenditures. $X_i = (X_{1i}, \ldots, X_{7i})'$ is a vector of covariates: $X_{1i} = 1$, X_{2i} is the logarithm of total expenditures, X_{3i} is log family size, X_{4i} age class, X_{5i} education level, X_{6i} is the degree of urbanization, and $X_{7i} = X_{2i}^2$.[2] $\alpha = (\alpha_1, \ldots, \alpha_7)'$ is

[2] Preliminary results with squares and cross-products of regressors suggest that only X_{2i}^2 should be incorporated.

an unknown parameter vector. Equation (1) can be derived from maximizing some quasi-concave household utility function, assuming absence of fixed costs of vacations. Since prices do not vary across the cross-section data, equation (1) is just an Engel curve, accounting for non-negativity. If $\alpha_7 = 0$, then the Engel curve corresponds to the Almost Ideal Demand System (Deaton and Muellbauer, 1980). Education level is included as a proxy for the cost of time, which might have a negative impact on vacation expenditures. Since the data do not contain information on wages, job level, or hours worked, time costs could not be taken into account in a more direct way.

ε_i is the error term. The models in this section make different assumptions about the conditional distribution of ε_i, given X_i. We assume throughout the paper that the (X_i, ε_i), $i = 1, \ldots, N$, are i.i.d.

We use a cross-section of $N = 1815$ families with at least two adults, taken from the Netherlands Consumer Expenditure Survey drawn in 1981. See Appendix 1 for details on the variables involved and sample statistics. We basically use the same data as Van Soest and Kooreman (1987).

Parametric Models

The common assumption about the distribution of the error terms in equation (1) is

$$\varepsilon_i \mid X_i \sim N(0, \sigma^2) \tag{2}$$

This yields the Tobit model (cf. Tobin, 1958, or Amemiya, 1973). Error terms are normal and independent of regressors. They are symmetric and homoscedastic.

ML estimates for the Tobit model and corresponding standard errors are mentioned in column I of Table I. The estimates for α_2 and α_7 imply that the budget share spent on vacations is increasing as a function of total expenditures in the whole sample range, implying that vacation is a luxury. The estimated *ceteris paribus* difference between $X_i'\alpha$ for families with total expenditures Dfl 50,000 and Dfl 20,000 is 0·052. This can be interpreted as the difference between budget shares spent on vacations, if ε_i and $X_{2i}, \ldots, X_{6i}$ are such that, for both families, vacation expenditures are positive. Young families spend significantly more on vacations than older families. Families in big cities spend more than those living in the country. The impact of family size and the family head's education level are insignificant.

The Tobit estimates will generally be inconsistent if the error terms are heteroscedastic or non-normal (see e.g. Hurd, 1979; Goldberger, 1983). These assumptions can be tested using Lagrange Multiplier tests, as in Chesher and Irish (1987). Both assumptions are clearly rejected at the 1% level, suggesting that at least one of the two assumptions is violated.

Explicit incorporation of parametric forms of heteroscedasticity is straightforward. We present results for the following two specifications:

$$\varepsilon_i \mid X_i \sim N(0, \exp(X_i'\beta)^2) \tag{3}$$

$$\varepsilon_i \mid X_i \sim N(0, \mid X_i'\Sigma X_i \mid), \quad \text{where } \Sigma = \mathrm{Diag}(\beta_1, \ldots, \beta_7) \tag{4}$$

In equation (3) heteroscedasticity is modelled in a multiplicative way. This has the advantage that the variance is guaranteed to be positive for all vectors β. Gabler *et al.* (1993) use equation (3) to test for heteroscedasticity in a binary choice model. From an economic point of view, equation (4) is more attractive than (3), if all components of β are non-negative. In this case, the latent model can be interpreted as a random coefficients model, with diagonal covariance

Table I. Estimation results parametric models (standard errors in parentheses)

	I	II	III	IV
α_1	−292·11 (94·83)	−515·21 (119·01)	−608·87 (125·64)	−492·87 (102·36)
α_2	50·43 (18·35)	92·37 (22·55)	110·18 (23·78)	88·49 (19·37)
α_3	−0·91 (0·56)	−0·38 (0·49)	−0·82 (0·38)	−0·39 (0·43)
α_4	0·22 (0·06)	0·14 (0·06)	0·14 (0·06)	0·14 (0·05)
α_5	0·27 (0·17)	0·34 (0·15)	0·29 (0·13)	0·32 (0·13)
α_6	0·56 (0·10)	0·39 (0·09)	0·43 (0·09)	0·39 (0·08)
α_7	−2·16 (0·88)	−4·12 (1·07)	−4·96 (1·14)	−3·95 (0·92)
σ	6·23 (0·10)			
β_1		18·46 (12·45)	1568·26 (269·99)	9·69 (14·02)
β_2		−2·86 (2·36)	−26·48 (4·69)	−1·21 (2·66)
β_3		−0·387 (0·065)	−8·23 (1·17)	−0·315 (0·078)
β_4		0·014 (0·006)	0·13 (0·04)	0·016 (0·008)
β_5		−0·047 (0·020)	−0·51 (0·14)	−0·035 (0·024)
β_6		0·044 (0·011)	0·29 (0·10)	0·045 (0·014)
β_7		0·124 (0·112)	0·11 (0·02)	0·044 (0·126)
γ_1				−0·061 (0·010)
log likelihood	−4233·86	−4155·45	−4157·73	−4144·15

The censored regression model (equation (1)) and regressors are described at the beginning of this section.
I: Tobit model (2)
II: Exponential heteroscedasticity and normality (3)
III: Random coefficients and normality (4)
IV: Exponential heteroscedasticity and non-normality (5)

matrix of the vector of coefficients. Equation (4) is used by Horowitz (1993) to allow for heteroscedasticity in a binary choice model.

Estimation results for the two models are mentioned in columns II and III of Table I. The signs of the estimates for the α_j are the same as in the Tobit model. Differences in magnitude and significance level with the Tobit model estimates are not very large either. The estimated differences between vacation budget shares of families with total expenditures Dfl 50,000 and Dfl 20,000, are 0·064 and 0·068, for models (3) and (4), respectively.

In both models, most of the heteroscedasticity parameters are significantly different from 0 on a 5% level. In the random coefficients model, three of them are negative, contradicting the random coefficients interpretation. For two observations, the estimate of $X_i'\Sigma X_i$ is negative, and the absolute value in model (4) is meaningful. According to likelihood ratio tests, the Tobit model is strongly rejected against both more general models. The likelihood of the exponential heteroscedasticity model is slightly better than that of the random coefficients model.

Both heteroscedasticity and non-normality can be incorporated into a parametric framework, using, for example, the following family of distributions for $\varepsilon_i \mid X_i$:

$$P[\varepsilon_i < t \mid X_i, \beta] = F(t \mid X_i, \beta) = G(t/h(X_i, \beta)) \tag{5}$$

with

$$G(s) = \Phi(\gamma_0 + s + \gamma_1 s^2 + \gamma_2 s^3) \tag{6}$$

Here $h(X_i, \beta)$ is a twice differentiable non-negative function, for example $h(X_i, \beta) = \exp(X_i'\beta)$. Φ denotes the standard normal distribution function. The distribution function G was proposed by Ruud (1984) as a family of probability distributions generalizing the standard normal. If $\gamma = (\gamma_0, \gamma_1, \gamma_2)' = 0$

then the conditional distribution of ε_i is normal with zero mean and standard deviation $h(X_i, \beta)$. In general, γ must be such that the corresponding density is everywhere non-negative, i.e.:

$$\gamma_2 \geq \gamma_1^2/3 \tag{7}$$

An identifying location restriction on the conditional distribution of ε_i (or $\varepsilon_i/h(X_i, \beta)$) is necessary, implying that γ_0 can be written as a function of γ_1 and γ_2. Natural choices are zero conditional median ($\gamma_0 = 0$) or zero conditional mean.

In Appendix 2 we show that the Chesher and Irish (1987) test for non-normality in the Tobit model can be generalized to a test for non-normality in models allowing for heteroskedasticity, by nesting the normal heteroscedastic model in the general model satisfying models (5) and (6). Tests derived for the zero conditional median and the zero conditional mean case are slightly different. All the tests lead to rejection of the null hypothesis, suggesting that, even if heteroscedasticity is accounted for, non-normality remains a problem. We therefore have also estimated the general models (5) and (6) by ML, allowing for multiplicative heteroscedasticity and non-normality. Results are in column IV of Table I. Restriction (7) appeared to be binding. Presented results are those with $\gamma_2 = \gamma_1^2/3$ imposed. A standard t-test and a likelihood ratio test result in rejecting normality, confirming the LM test results. However, the estimates for α and β, and their standard errors, correspond surprisingly well to those in column II of the table.

The implied *ceteris paribus* pattern of vacation expenditures as a function of total expenditures is similar to our previous findings: The maximum budget share is obtained for total expenditures Dfl 73,000. The difference between the predicted shares if total expenditures are Dfl 50,000 and Dfl 20,000, is 0·061.

Finally, we carried out several chi-square goodness of fit tests for the parametric models (cf. Andrews, 1989). These tests are based upon classifying observations into cells and comparing estimated cell probabilities (conditional on the covariates) with sample probabilities. See Appendix 2 for details. In all cases, the hypothesis of no misspecification was rejected.

Semi-parametric Models

Various semi-parametric estimators for the censored regression model (1) are available. They require different assumptions on the conditional distribution of ε_i for given X_i for consistency. We consider estimators characterized by the weak identifying restriction that ε_i has zero median:

$$\text{Med}(\varepsilon_i \mid X_i) = 0 \tag{8}$$

Two such estimators are available. The first is Censored Least Absolute Deviations (CLAD) (Powell, 1984). CLAD minimizes the average absolute deviation between y_i and $\max\{0, X_i'\alpha\}$. It can be interpreted as a conditional median estimator, since model (8) implies that $\text{Med}(y_i \mid X_i) = \text{Max}\{0, X_i'\alpha\}$. CLAD is consistent and asymptotically normal under mild regularity conditions. Powell (1984) derives a consistent estimator for its asymptotic covariance matrix. CLAD does not attain the asymptotic efficiency bound.

Second, we consider the estimator of Newey and Powell (1990). The underlying idea is to apply optimally weighted CLAD. Optimal weights are derived from the efficient score, and estimated non-parametrically, using a consistent first-step estimator. For technical reasons, sample splitting is used, so that the first step consists of two CLAD estimations for two disjoint subsamples. The second step consists of one Newton–Raphson step in the direction of minimizing the optimally weighted sum of least absolute deviations (Newey and Powell, 1990, equation (4.9)). Under some regularity conditions, this estimator is semi-parametrically efficient. A consistent estimator of its asymptotic covariance matrix is available.

CLAD requires minimization of a nondifferentiable function. For this purpose, we used the

simplex algorithm introduced by Nelder and Mead (1965) and extended by O'Neill (1971). Estimation results are mentioned in Table II. Estimating the covariance matrix of the CLAD estimator involves choosing smoothing parameters (c_0 and γ in Powell, 1984, equations (5.5) and (5.6)). Following Powell, we chose $\gamma = 0{\cdot}2$ and tried various values of c_0.[3] We present results for three values of c_0. The choice of c_0 appears to have a small impact only.

Table II(a). Estimation results CLAD (standard errors in parentheses)

	I	II	III
α_1	−731·30 (180·65)	(258·74)	(211·13)
α_2	133·99 (33·68)	(48·15)	(39·57)
α_3	−0·23 (0·90)	(0·78)	(0·74)
α_4	0·14 (0·10)	(0·11)	(0·09)
α_5	0·14 (0·18)	(0·18)	(0·22)
α_6	0·49 (0·11)	(0·11)	(0·12)
α_7	−6·12 (1·57)	(2·23)	(1·85)

The model assumptions are given by equations (1) and (8). The regressors are described at the beginning of Section 2.
I: CLAD estimator and standard errors with smoothness parameters $c_0 = 0{\cdot}74$ and $\gamma = 0{\cdot}20$.
II: CLAD standard errors with smoothness parameters $c_0 = 0{\cdot}35$ and $\gamma = 0{\cdot}20$.
III: CLAD standard errors with smoothness parameters $c_0 = 1{\cdot}5$ and $\gamma = 0{\cdot}20$.

Table II(b). Estimation results Newey–Powell estimator (standard errors in parentheses)

	I	II	III
α_1	−723·69 (99·01)	−682·42 (71·44)	−680·34 (127·42)
α_2	132·65 (18·70)	124·64 (13·54)	124·43 (24·01)
α_3	−0·67 (0·28)	−0·67 (0·20)	−0·63 (0·37)
α_4	0·17 (0·04)	0·14 (0·02)	0·17 (0·05)
α_5	0·13 (0·09)	0·12 (0·07)	0·08 (0·11)
α_6	0·47 (0·06)	0·43 (0·05)	0·49 (0·07)
α_7	−6·07 (0·88)	−5·67 (0·64)	−5·67 (1·13)
		IV	V
α_1		−733·89 (84·78)	−791·72 (123·15)
α_2		134·70 (16·07)	145·16 (23·24)
α_3		−0·43 (0·25)	−0·67 (0·34)
α_4		0·13 (0·03)	0·16 (0·04)
α_5		0·17 (0·08)	0·16 (0·11)
α_6		0·48 (0·05)	0·50 (0·07)
α_7		−6·16 (0·76)	−6·63 (1·10)

The model assumptions are given by equations (1) and (8). The regressors are described at the beginning of Section 2.
I: Newey–Powell estimator with smoothness parameters $h = 0{\cdot}22$ and $k = 40$.
II: Idem with $h = 0{\cdot}11$, $k = 40$; III: idem with $h = 0{\cdot}44$, $k = 40$; IV: idem with $h = 0{\cdot}22$, $k = 20$; V: idem with $h = 0{\cdot}22$, $k = 80$.

[3] For this problem, no rules for bandwidth choice are available. Instead, we used a rule of thumb for bandwidth choice in kernel density estimation to get some idea about reasonable values of c_0, since the situation there is similar.

For the Newey–Powell estimator, the choice of smoothness parameters (Newey and Powell, 1990, equation 4.5) affect not only the covariance matrix but also the estimates themselves. We present results for five choices of smoothing parameters. Results are somewhat sensitive to the choice. In particular, this applies to standard errors and inference. For instance, a Hausman-type specification test comparing outcomes of CLAD and Newey–Powell estimates results either in clear rejection or clear non-rejection of the hypothesis of no misspecification, depending upon the specific values of the smoothness parameters. However, most of the parameter estimates according to Newey–Powell and CLAD seem to be quite similar.

According to the reported estimates in columns I of Tables II(a) and II(b), the budget share spent on vacations is maximal if total expenditures are Dfl 57,000 (CLAD) or 56,000 (Newey–Powell). The difference between the shares for families with total expenditures Dfl 50,000 and Dfl 20,000 is 0·066 (CLAD) or 0·064 (Newey–Powell). These numbers are not too much out of line with their parametric counterparts. This is also the case for estimated coefficients of the other regressors.

From a policy point of view, the most interesting feature of the results is the income elasticity of aggregate vacation expenditures. This is, for example, what is used in a complete model of the Dutch tourist market (SEO, 1991). In the semi-parametric model we do not know the distribution of the error terms. All we can compute is the percentage change in the sample average of the conditional median of vacation expenditures if income of all families rises by 1%.

Ninety per cent confidence intervals for these elasticities are mentioned in Table III.[4] The results strongly suggest that vacations are a luxury: the lower bounds of the confidence intervals exceed one and, with one exception, are larger than two. All intervals overlap. The parametric elasticity estimates are slightly larger than the semi-parametric ones. The confidence intervals according to the (efficient) Newey–Powell estimator are substantially smaller than those of the (inefficient) CLAD estimator. The choice of smoothness parameter has some effect; the length of the largest and smallest interval according to the Newey–Powell estimates differ by a factor of 1·6.

Table III. Elasticities censored regression models; 90% confidence bounds

	Lower bound	Upper bound
Parametric models:		
Tobic model	2·20	2·87
Exp heteroscedastic	2·25	2·87
Exp. heteroscedastic/non-normal	2·26	2·76
Semi-parametric models:		
CLAD ($c_0 = 0{\cdot}74$, $\gamma = 0{\cdot}20$)	2·04	2·52
CLAD ($c_0 = 0{\cdot}35$, $\gamma = 0{\cdot}20$)	1·82	2·56
CLAD ($c_0 = 1{\cdot}5$, $\gamma = 0{\cdot}20$)	2·00	2·57
Newey–Powell ($h = 0{\cdot}22$, $k = 40$)	2·15	2·41
Newey–Powell ($h = 0{\cdot}11$, $k = 40$)	2·27	2·49
Newey–Powell ($h = 0{\cdot}44$, $k = 40$)	2·13	2·48
Newey–Powell ($h = 0{\cdot}22$, $k = 20$)	2·12	2·39
Newey–Powell ($h = 0{\cdot}22$, $k = 80$)	2·18	2·53

[4] The elasticities are calculated 500 times, for 500 independent draws of the parameters from the estimated asymptotic distribution of the estimator. For each case, we present the 0·05 and 0·95 quantiles.

Various other semi-parametric estimators for the censored regression model (1) are available. They require other assumptions on the conditional distribution of ε_i given X_i for consistency. Many of them require independence between theerror term ε_i and the covariates X_i (Duncan, 1986; Fernandez, 1986; Horowitz, 1986; Ruud, 1986). Powell's (1986b) symmetrically trimmed least squares estimator requires the conditional distribution of the errors to be symmetric.

We tested the assumptions of independence and symmetry following Powell (1986a): CLAD can easily be generalized, replacing model (8) by another conditional quantile restriction. We estimated the models based upon the restrictions of 25%, 37·5%, 62·5%, and 75% quantiles. If errors and regressors are independent, the estimates of the slope parameters should be similar in each case. Using a generalized Hausman test based upon the differences, the null of independence is clearly rejected. Under conditional symmetry, the differences between p quantile parameters and median parameters must be opposite to those between median and $(1-p)$ quantile parameters, including the constant term (cf. Powell, 1986a, p. 155). The hypothesis is rejected using the 37·5% and 62·5% quantiles. These test results correspond to the findings based on parametric models.

As a consequence of these rejections, we did not consider it worthwhile to consider the alternative estimators mentioned above. Neither did we consider linear combinations of various quantile estimators, which, in case of independence, might be more efficient than a single one such as CLAD.

3. A TWO-EQUATIONS MODEL

The results of the Hausman test comparing CLAD and Newey–Powell estimates in Section 2 suggest that the semi-parametric censored regression model is still misspecified. The single index specification might be too restrictive. A possible explanation may be the fact that the data contain not only many zero expenditures observations but also relatively few observations with small positive vacation expenditures (cf. Table A.II in Appendix 1). This is not captured in the censored regression model. In this section we consider a model in which the decisions on whether to spend anything or not, and on how much to spend, are separated:

$$P(a_i = 1 \mid X_i) = F(X'_{ai}\alpha_a); \quad P(a_i = 0 \mid X_i) = 1 - F(X'_{ai}\alpha_a) \tag{9}$$

$$y_i^* = X'_{bi}\alpha_b + \varepsilon_i \tag{10}$$

$$y_i = 0 \text{ if } a_i = 0; \quad y_i = y_i^* \text{ if } a_i = 1 \tag{11}$$

Here $a_i = 1$ indicates participation ($y_i > 0$), $a_i = 0$ indicates non-participation ($y_i = 0$). X_i is a vector of regressors, X_{ai} and X_{bi} are subvectors of X_i. $F: \mathbb{R} \to [0, 1]$ is unknown (and not necessarily a distribution function). Equation (9) is a single index binary choice equation: X_i only enters through the index $X'_{ai}\alpha_a$. To identify α_a, some normalization must be added.

The regression equation (10) explains the budget share spent on vacation, conditional upon the decision to go on holiday. We cannot think of economic arguments for excluding regressors from either equations (9) or (10). Therefore, in principle, equations (10) and (11) may include the same regressors. Instead of imposing exclusion restrictions on the regressors in equation (10) we make the identifying assumption

$$E(\varepsilon_i \mid X_i, a_i = 1) = 0 \tag{12}$$

This assumption makes it possible to estimate equation (10) separately from equation (9), using observations with $y_i > 0$ only. Without equation (12), and without exclusion restrictions on the

regressors, identification of the model relies on the linearity of the systematic part of equation (10). A test of equation (12), to be discussed below, may therefore not be very powerful.

Note the difference in interpretation between the common selectivity model (Tobit II in Amemiya, 1984) and the model here. The standard example of the latter is a wage equation combined with a binary choice employment equation. In this model, the wage rate for someone who does not work has a clear interpretation: potential earnings if he or she would find a job. In our case, however, vacation expenditures of non-participants are zero by definition. The problem is not modelling selectivity but zero expenditures. Whereas a continuous distribution of (positive) potential wage rates in the population of workers and non-workers makes sense, the concept of positive potential vacation expenditures of people who do not spend anything does not seem very useful.

In the remainder of this section, we analyse first equation (9) and then equation (10).

The Binary Choice Participation Equation

We first analyse the Probit model, arising if the unknown F in equation (9) is specified as

$$F(z) = \Phi(z/\sigma_a) \tag{13}$$

for some $\sigma_a > 0$. Φ is the standard normal distribution function. In the vector of covariates X_{ai} we included $X_{1i}, \ldots, X_{6i}$ (cf. Section 2). On the basis of preliminary estimation results with squares and cross-products of the regressors we also included the cross-term $X_{4i}X_{5i}$. In contrast to Section 2, we found no reason to include X_{2i}^2.

ML estimates of the Probit model are presented in column I of Table IV. To be able to compare Probit with semi-parametric results discussed below, we have normalized α_{a2}, the slope coefficient of log(total expenditures), to 1. According to the Probit estimates, the probability of going on vacation increases with family size and degree of urbanization, but only the latter is significant. The significant cross-term between age and education level may reflect different life cycle patterns of vacation behaviour of the low and high educated. The former tend to participate at a later stage.

Since the ML Probit estimator for α_a may be inconsistent if the normality assumption (13) is incorrect, we performed various specification tests. Hardly any misspecification is detected.

Table IV. Estimation results of the binary choice model (standard errors in parentheses)

Parameter	I	II	III
α_{a1} (constant term)	−10·896 (0·227)	0	0
α_{a2} (log total expend.)	1	1	1
α_{a3} (log family size)	0·050 (0·083)	0·052 (0·067)	0·228 (0·113)
α_{a4} (age class)	0·060 (0·019)	0·093 (0·012)	0·080 (0·022)
α_{a5} (education level)	0·180 (0·063)	0·236 (0·043)	0·333 (0·069)
α_{a6} (degree of urbanization)	0·045 (0·015)	0·051 (0·012)	0·072 (0·020)
α_{a7} ($X_{4i}X_{5i}$)	−0·017 (0·007)	−0·023 (0·005)	−0·016 (0·008)
σ_a^2	0·606 (0·104)		
log L	−1082·3	−1090·4	−1098·3

The first six regressors are the same as those in Section 2; $X_{7i} = X_{4i}X_{5i}$.
I: ML estimates Probit model (1), (13) (normalization: $\alpha_{a2} = 1$).
II: Klein–Spady estimates (9) (normalization: $\alpha_{a1} = 0$, $\alpha_{a2} = 1$; smoothing parameter $h_N = 0{\cdot}2$).
III: Klein–Spady estimates (g) ($\alpha_{a1} = 0$, $\alpha_{a2} = 1$; smoothing parameter $h_N = 0{\cdot}4$).
log L: log-likelihood value for Probit; quasi-log-likelihood value (no trimming) for Klein–Spady.

Neither normality nor homoscedasticity were rejected by the score tests of Chesher and Irish (1987). Chi-square tests (Andrews, 1989) did not result in rejecting the Probit model, with one exception at the 5% level, and no exceptions at the 1% level.

The problem with these chi-square diagnostics is that results may strongly depend on the arbitrary choice of partitioning the sample space. Therefore, we also tested equation (13) using a test of Horowitz (1993, proposition 1). This test is based on a non-parametric regression of a_i on $X'_{ai}\hat{\alpha}_i/\hat{\sigma}_a$ and a uniform confidence band for the regression function. The Probit model is rejected if the standard normal distribution function does not lie within this band. The outcome of the test is presented in Figure 1, which includes a plot of Φ, the uniform 95% confidence band, and the values of the kernel regression in all points $X'_{ai}\hat{\alpha}_a/\hat{\sigma}_a$, $i = 1, \ldots, N$. Since Φ lies entirely within the 95% confidence band, equation (13) is not rejected at the 5% level.

An asymptotically efficient semiparametric estimator for equation (9) is given by Klein and Spady (1993). It is essentially a quasi maximum likelihood estimator: The unknown function F in the likelihood is replaced by a non-parametric Kernel regression estimate of a_i on $X'_{ai}\alpha_a$. The quasi-loglikelihood is then maximized with respect to α_a. Klein and Spady (1993) show that the resulting estimator of α_a is consistent, asymptotically normal, and asymptotically efficient. We used the following bias reducing kernel (cf. Klein and Spady's assumption C.8a):

$$K(z) = (3/2 - (1/2)z^2)\phi(z) \tag{14}$$

with ϕ the standard normal density. Following the Monte Carlo study in Klein and Spady (1993), we present the outcomes without probability or likelihood trimming.

To guarantee identification of α_a in equation (9) two normalizations are necessary: the constant term is set equal to zero, since it is absorbed in F. The coefficient of X_{2i} (log total expenditure, the only continuous variable in X_{ai}), is set equal to one. Klein–Spady estimates are in columns II and III of Table IV. The outcomes are somewhat sensitive to the choice of the

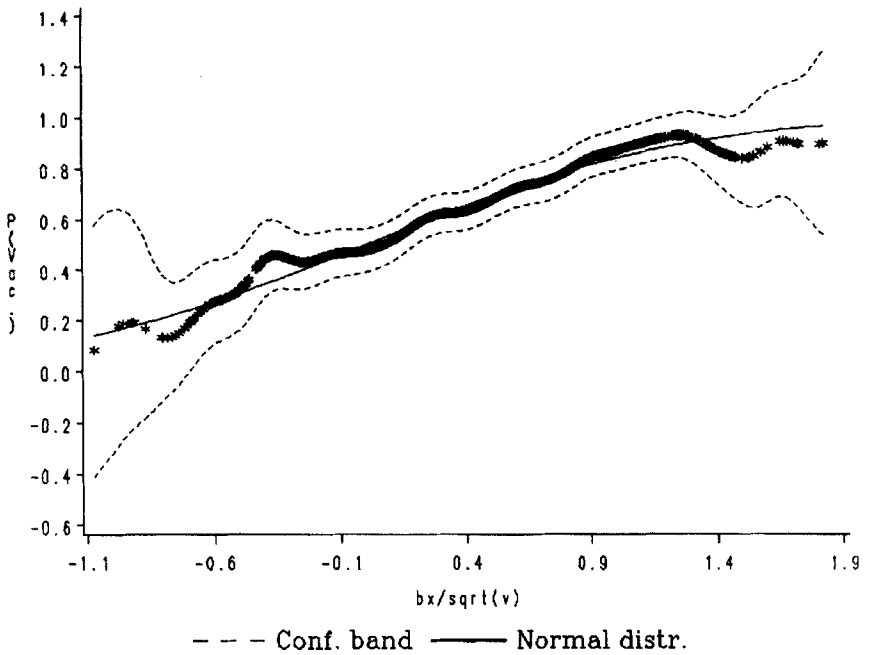

Figure 1. P(Vac.) from Probit and non-parametric regression

smoothness parameter h_N in the kernel. We present results for $h_N = 0{\cdot}2$ and $h_N = 0{\cdot}4$. Klein–Spady estimates are close to Probit estimates, particularly for $h_N = 0{\cdot}2$: The mean function increases with family size and the degree of urbanization; it increases and decreases with age for low and high education levels, respectively.

The function F_N, the estimate of F (cf. equation (23) in Klein and Spady, 1993; $h_N = 0{\cdot}2$) is presented in Figure 2, together with the function $G(z) = \Phi((z - 10{\cdot}90)/0{\cdot}77)$, the corresponding function for Probit. The Klein–Spady estimate F_N is initially steeper than its Probit counterpart, and decreases for high values of the mean function. Notable differences only occur in the region where observations are sparse, however. This is in line with conclusions from Figure 1.

In conclusion, the results suggest that the Probit specification works quite well for the data at hand. Not surprisingly, the single index model, of which Probit is a special case, yields quite similar results and also works quite well. This is also confirmed by results of informal graphical tests suggested by Horowitz (1993, pp. 61–62), which compare predicted and actual cell probabilities.[5]

The Regression Equation

We estimate equation (10) with and without imposing equation (12). Relaxing equation (12) can be achieved in a natural way by assuming

$$E(\varepsilon_i \mid X_i, a_i = 1) = \mu(X'_{ai}\alpha_a) \tag{15}$$

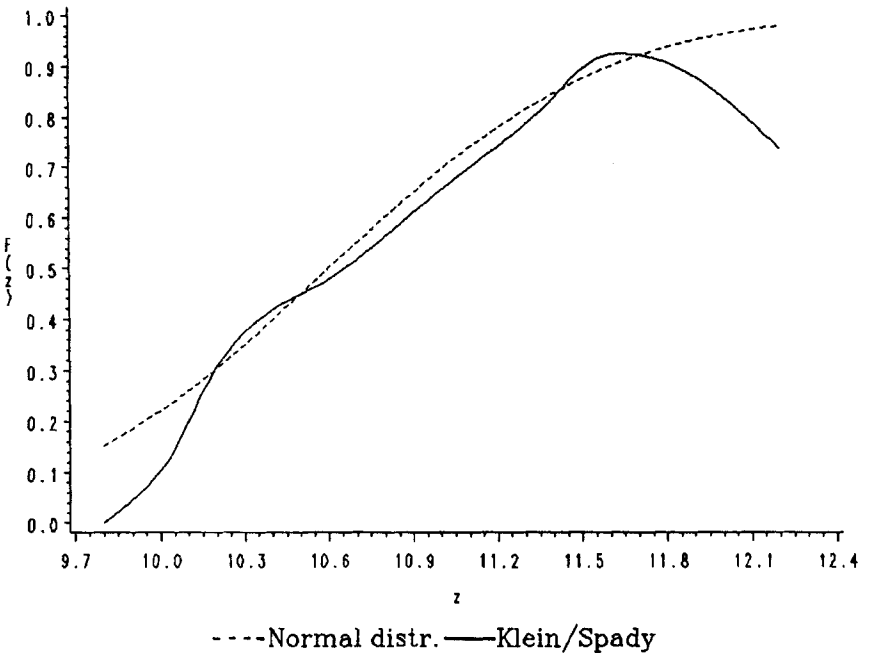

Figure 2. Estimated distribution functions

[5] Details for the test results of Probit as well as Klein–Spady are available from the authors upon request. Alternatively, the Probit specification could also be tested by a Wu–Durbin–Hausman style test, based upon differences between ML Probit and Klein–Spady estimates.

where $\mu: \mathbb{R} \to \mathbb{R}$ is unknown. This corresponds to the assumptions for the selectivity model of, for example, Newey (1988) or Powell (1987). The unknown function μ can now be replaced by the series approximation

$$\mu(z) \approx \mu_0 + \mu_1 \lambda(z) + \mu_2 \lambda^2(z) + \cdots + \mu_K \lambda^K(z) \tag{16}$$

where $\lambda(z)$ is the inverse Mill's ratio. This approximation has the advantage of nesting the case of bivariate normality ($K = 1$). μ_0 will be absorbed in the constant term α_{b0}.

The vector of covariates X_{bi} includes $X_{1i}, \ldots, X_{6i}$, described in Section 2, and $X_{7i} = X_{2i}X_{6i}$. The latter is the only cross-product retained after preliminary estimations with squares and cross products.

In Table V, we present the results of three estimators. The first two are based upon equations (10) and (12). First, α_b is estimated by OLS of y_i on X_{bi}, using the subsample $\{i; y_i > 0\}$ (column I). Standard errors are computed in two ways: first in the standard way, assuming errors ε_i to be homoscedastic and independent of X_i. The second set of standard errors is based upon an estimator of the covariance matrix of the OLS-estimator which, under weak regularity conditions, remains consistent in case of heteroscedasticity.[6] Differences between the two are quite small. Results imply that the *ceteris paribus* effect of log expenditures is large and positive for people living in the country but small or even negative for people living in cities. There are some notable differences between effects on participation and effects on budget shares, conditional on participation. For example, family size enters with insignificant positive sign in the participation equation, but with a significant negative sign in the budget share equation. These differences suggest that using two separate equations instead of one censored regression equation is useful. They may, for example, point at fixed costs of vacations, which in equations (9)–(11) are incorporated in the reduced form.

OLS is asymptotically efficient if ε_i is independent of X_i and normally distributed. These assumptions were tested using the truncated sample. Normality was strongly rejected by a score test. Tests against heteroscedasticity similar to those of Section 2 lead to a clear rejection of the null of homoscedastic error terms.

These results suggest that normality or homoscedasticity are violated. In this case, retaining equation (12), OLS is consistent, but not efficient. An asymptotically efficient semiparametric estimator of α_b for this case is given by Robinson (1987). It is a weighted least squares estimator, where the weights are non-parametric estimates of $V\{\varepsilon_i \mid X_{bi}\}^{-1}$. Weights are constructed from OLS residuals, using a k-nearest-neighbour (NN) estimator with uniform weights (Robinson, 1987, p. 879). Estimation results for $k = 150$ are in column II. Changing the smoothness parameter k hardly affected the outcomes. The estimates are quite close to the OLS estimates. Estimated standard errors of the Robinson estimator are smaller than (heteroscedasticity corrected) OLS standard errors, but the differences are quite small.

Differences between asymptotically efficient Robinson estimates and consistent OLS estimates can be used to construct a Hausman-type specification test for equations (10) and (12). This does not lead to rejection of the nullhypothesis of no misspecification. Note, however, that the power of this test may not be very large.

The estimates in column III account for possible selectivity. Equation (12) is replaced by equation (15) and (16). Powers of $\lambda(X'_{ai}\hat{\alpha}_a)$, where $\hat{\alpha}_a$ is the ML Probit estimate of α_a, are added to the regressors in equation (10). The resulting regression equation is then estimated using the

[6] These (Eicker–White-type) standard errors are obtained from the following consistent estimator of the OLS covariance matrix: $\hat{V} = [N^{-1}\Sigma_i X'_{bi}X'_{bi}]^{-1}[N^{-1}\Sigma_i X'_{bi}X'_{bi}\hat{\sigma}_i^2][N^{-1}\Sigma_i X'_{bi}X'_{bi}]^{-1}$, where $\hat{\sigma}_i^2$ is the non-parametric estimator of $V\{\varepsilon_i \mid X_{bi}\}$ used in the Robinson estimator (Robinson, 1987, p. 879).

Table V. Estimation results of the regression model

Parameter	I	II	II
α_{b1} (constant)	−22·00	−18·67	19·91
	(10·04)	(8·54)	(25·05)
	[9·21]		
α_{b2} (log total exp)	2·57	2·28	−1·04
	(0·97)	(0·83)	(2·09)
	[0·89]		
α_{b3} (log family size)	−1·46	−1·56	−1·63
	(0·42)	(0·37)	(0·41)
	[0·39]		
α_{b4} (age class)	0·15	0·15	−0·07
	(0·05)	(0·04)	(0·16)
	[0·05]		
α_{b5} (education)	0·002	0·02	−0·63
	(0·13)	(0·11)	(0·45)
	[0·12]		
α_{b6} (urbanization)	5·37	4·28	0·31
	(2·25)	(2·12)	(0·11)
	[2·28]		
α_{b7} $(X_{2i}X_{6i})$	−0·47	−0·37	0·08
	(0·22)	(0·20)	(0·05)
	[0·22]		
λ	—	—	−2·85
			(3·18)

The model specification is given in equation (16). The first six regressors are the same as in Section 2; $X_{7i} = X_{2i}X_{6i}$; λ is the coefficient corresponding to the inverse of Mill's ratio.
I: Ordinary Least Squares estimates; (.): standard errors, computed in the standard way, [.]: heteroscedasticity corrected standard errors.
II: Robinson estimates; smoothness parameter $k = 150$, (.): standard errors.
III: Robinson estimates with the inverse of Mill's ratio included (without cross-term), (.): standard errors.

Robinson estimator. Estimates for $K = 1$ are in column III of Table V. As to be expected, the estimates are very inaccurate, due to multicollinearity between $\lambda(X'_{ai}\hat{\alpha}_a)$ and X_{bi}, even though X_{ai} and X_{bi} contain a different cross-product.[7] Under the null hypothesis that equation (12) holds, standard errors do not have to be corrected for replacing α_a by its estimate. According to the t-value of $\lambda(X'_{ai}\hat{\alpha}_a)$, this null hypothesis is not rejected. We conclude that equations (10) and (12) describe the truncated sample appropriately. As we have argued at the outset of this section, it may be the case that not finding selectivity effects is due to identification problems.

4. EVALUATION AND CONCLUSIONS

We have analysed a number of models to explain vacation expenditures of Dutch families. In Section 2 we considered the single-equation censored regression model. All parametric

[7] Results are more accurate if the cross-product is excluded from X_{bi}, but $\lambda(X'_{ai}\hat{\alpha}_a)$ remains insignificant (with t-value 0·2). We also tried higher-order polynomials of λ, but significance levels remained very low.

specifications we considered are rejected by specification tests. We estimated a semi-parametric specification based upon a weak zero-median assumption, using a consistent and an efficient estimator. It appeared that resulting estimates were sensitive to chosen smoothness parameters, in particular the standard errors and inference based upon them. For instance, the result of a Hausman type test for model misspecification depends on the chosen bandwidth.

In Section 3 we considered models in which the decisions whether or not to go on holiday and how much to spend are modelled separately. The participation decision was modelled by a semi-parametric single index specification and by a simple Probit model. Both appeared to fit the data quite well. Predicted probabilities according to these models differ for sparse values of the covariates only.

Conditional on the decision to participate, the decision on how much to spend was modelled by a regression equation. This equation was estimated using both OLS and a semi-parametrically efficient estimator, with almost identical results. Independence between the error in the regression equation and the error in the Probit model was tested and not rejected.

Even if selectivity is allowed for, the general one-equation model is not nested in the two-equations model. For example, if the errors in model (1) are heteroscedastic, the participation probability $P\{y_i^* > 0 \mid X_i\}$ of the one-equation model will not satisfy the single index restriction (9). Formal tests of the one-equation model against the two-equation model could therefore not be performed.

The two-equation model has several advantages over the censored regression model. It is more flexible, in the sense that participation and budget share level decision are separated. Fixed costs of vacations are thus allowed for. According to the estimated slope coefficients, this indeed leads to some improvement. It also has the advantage that, even for a semi-parametric specification, it can be used to compute expected vacation expenditures for each family, using

$$E(y_i \mid X_i) = E(y_i \mid a_i = 1, X_i)P(a_i = 1 \mid X_i) \tag{17}$$

It thus also allows for the computation of, for example, the elasticity of aggregate vacation expenditures with respect to total expenditures. We find aggregate income elasticities of about 0·7 for the participation probability, and 1·7 for expected vacation expenditures. Such calculations are not possible in the semiparametric censored regression model. For the parametric censored regression models in Section 2, we find similar income elasticities. On the other hand, the models in Section 3 capture the average value of vacation expenditures in the data much better than those in Section 2. This is one more reason to prefer the two-equation model.

Semi-parametric estimators are at this moment only available for some specific, relatively simple models. The recursive model in Section 3 serves to illustrate that such simple models can be combined into a model which in some sense captures the complicated economic phenomena of interest, although it is clear that this cannot be done without sacrificing some of the economic structure.

APPENDIX 1: THE DATA

The data originate from the Consumer Expenditure Survey drawn in 1981 by the Netherlands Central Bureau of Statistics, now named Statistics Netherlands. Sample statistics for the exogenous variable are given in Table A.I.

Vacation expenditures are defined as expenditures of any member of the family on a vacation. A vacation is defined as a 'stay away from home for recreation purposes for at least four successive nights'. Vacation expenditures are zero for 37% of all families in the sample. The average annual amount spent on vacations per family is Dfl 1415·4, zeroes included. The distribution of positive budget share of vacation expenditures is presented in Table A.II.

Table A.I. Sample statistics

	All observations		Observations with zero vacation expenditures		Observations with non-zero vacation expenditures	
	Mean	std dev.	Mean	std dev.	Mean	std dev.
X_{2i}	10·37	0·36	10·22	0·34	10·46	0·35
X_{3i}	1·10	0·37	1·05	0·36	1·13	0·37
X_{4i}	7·09	3·18	7·32	3·44	6·96	3·01
X_{5i}	2·31	1·06	2·06	0·96	2·45	1·09
X_{6i}	3·86	1·74	3·76	1·77	3·91	1·72

X_{2i}: logarithm of total family expenditures in 1981 (in Dfl)
X_{3i}: logarithm of family size ($2 <$ family size < 7)
X_{4i}: age class family head; $X_{4i} = 1$: <20 years old; $X_{4i} = 2$: 20–24 years old; $X_{4i} = 3$: 25–29 years old; ...; $X_{4i} = 13$: >74 years old
X_{5i}: education level family head, ranging from 1 (low) to 5 (high)
X_{6i}: degree of urbanization, ranging from 1 (country village) to 6 (large city)

Table A.II. Frequency distribution of budget shares spent on vacations: positive shares only, 1143 observations

Share (%)	Frequency (%)	Share (%)	Frequency (%)
0–1	4·2	8–9	6·0
1–2	8·5	9–10	3·9
2–3	10·7	10–11	3·3
3–4	11·8	11–12	2·7
5–6	12·7	12–13	1·4
6–7	8·0	13–14	1·0
7–8	8·5	>14	4·2

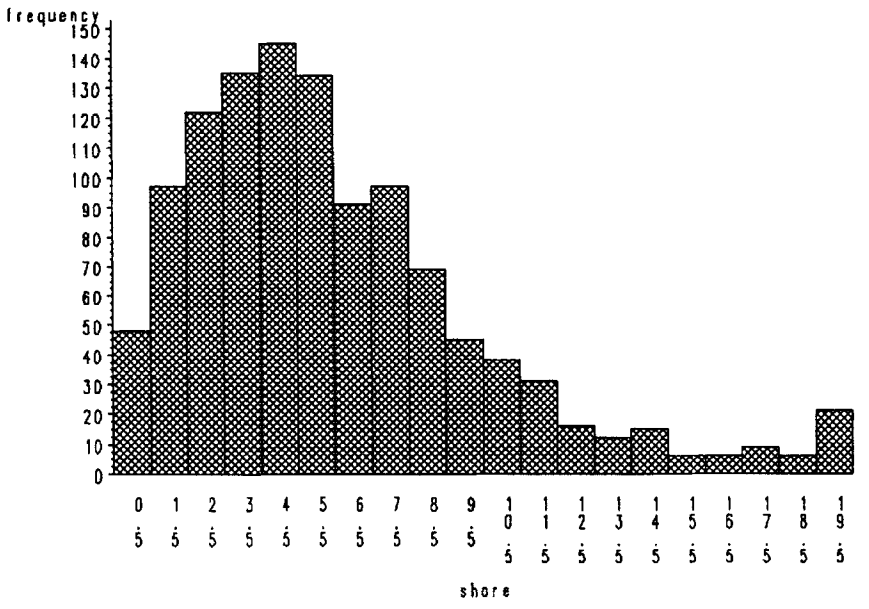

Figure A1. Distribution of budget shares spent on vacations (in %)

APPENDIX 2: SPECIFICATION TESTS

We present some details of the specification tests for the parametric censored regression model in Section 2.

Score Test for Non-normality in Model with Heteroscedasticity

We derive tests against non-normality in the model given by equations (1) and (3), using a score test based upon nesting (1) and (3) in (1) and (5) and (6). We impose either a zero mean or a zero median restriction on the conditional distribution of ε_i:

(a) $F(0 \mid X_i, \beta) = 1/2$ (zero conditional median): $\gamma_0 = 0$
(b) $E(\varepsilon_i \mid X_i, \beta) = 0$ (zero conditional mean): $\psi_0 = \psi(\gamma_1, \gamma_2)$, where ψ satisfies

$$\psi(0,0) = 0; \quad \partial\psi(0,0)/\partial\gamma_1 = -1; \quad \partial\psi(0,0)/\partial\gamma_2 = 0 \tag{A.1}$$

Conditions (9) are verified by substituting $\psi(\gamma_1, \gamma_2)$ for γ_0 and differentiating the equality $\int_R s \, dG(s) = 0$ with respect to γ_1 and γ_2. In the homoscedastic case, (a) and (b) are equivalent and necessary and sufficient to identify the constant term α_1 in equation (1). In case of heteroscedasticity, the equivalence no longer holds. Log-likelihood contributions of observations (y_i, X_i) are as follows:

$$\text{If } y_i = 0\text{: } \log F(-X_i'\alpha \mid X_i, \beta) = \log G(s_i), \quad \text{with} \quad s_i = -X_i'\alpha / h(X_i, \beta)$$
$$\text{If } y_i > 0\text{: } \log f(y_i - X_i'\alpha \mid X_i, \beta) = -\log h(X_i'\beta) + \log g(s_i), \quad \text{with} \quad s_i = (y_i - X_i'\alpha)/h(X_i, \beta) \tag{A.2}$$

The Chesher and Irish (1987) test for non-normality in the Tobit model can be interpreted as a score test on $(\gamma_1, \gamma_2)' = (0,0)'$ for the homoscedastic case of equations (5) and (6). To obtain a test for non-normality which remains appropriate in case of a parametric form of heteroscedasticity, one estimates equations (5) and (6) by ML, without imposing homoscedasticity but imposing normality, and then performs a score test on $(\gamma_1, \gamma_2)' = (0,0)'$. The test statistic is easily derived and, as in Chesher and Irish (1987), can be rewritten in terms of the generalized residuals

$$e_i^{(k)} = E_{\gamma=0}\{(\varepsilon_i / h(X_i'\beta))^k \mid y_i, X_i, \beta\} - \mu(k)$$

where $\mu(k) = \Gamma(k)/\{2^{k/2-1}\Gamma(k/2)\}$. The result depends on the restriction imposed, zero conditional median (case (a)) or zero conditional mean (case(b)). In either case, the test statistic can be obtained as the explained sum of squares in a regression of a vector $(1, \ldots, 1)' \in \mathbb{R}$ on the columns of an $N \times m$ matrix, where N is the number of observations and m is 2 plus the number of free parameters in α and β. If $h(X_i, \beta) = \exp(X_i'\beta)$, the typical row entries in this matrix are $e_i^{(1)} X_i / \exp(X_i'\beta)$, $e_i^{(2)} X_{ji}$, $j = 1, \ldots, 7$, either $-e_i^{(3)} + 2e_i^{(1)}$ (case (a)) or $-e_i^{(3)} + 3e_i^{(1)}$ (case (b)), and $-e_i^{(4)} + 3e_i^{(2)}$, where unknown parameters must be replaced by their ML estimates under the null ($\gamma = 0$). Under the null, the test statistic asymptotically follows a χ^2_2-distribution.

In case of homoscedasticity ($\beta_j = 0$ for $j > 1$) the two statistics for cases (a) and (b) coincide and are identical to the standard test statistic in Chesher and Irish (1987). In case of heteroscedasticity, however, variants (a) and (b) are not identical, since the transformed covariates $X_{ji}/\exp(X_i'\beta)$ will generally not contain a constant term.

The test results for $h(X_i, \beta) = \exp(X_i'\beta)$ are mentioned below. In order to take account of the inequality restriction (9), we also present the test results for one parameter restriction only:

H_0	Degrees of freedom	Test statistic test (a)	Test (b)	Critical value level 0·01
$\gamma_1 = \gamma_2 = 0$	2	13·1	13·0	9·2
$\gamma_1 = 0$	1	12·3	12·2	6·6
$\gamma_2 = 0$	1	1·0	1·0	6·6

The conclusion is clear: normality, in particular the restriction $\gamma_1 = 0$, is rejected at the usual significance levels. The difference between variants (a) and (b) is very small. Since the conditional distribution is symmetric if and only if $\gamma_1 = 0$, the results suggest that the distribution of the errors is skewed.

Chi-square Diagnostics

We performed several general specification tests suggested by Andrews (1989). These tests are based upon classifying observations into cells and comparing estimated cell probabilities (conditional on covariates) with sample probabilities. Andrews (1989) indicates how the cells can be chosen for the Tobit model and his strategy can easily be generalized to other parametric censored regression models. We used a partitioning based upon the endogenous variable into five cells: $y_i = 0$, and four cells with $y_i > 0$, distinguished by the value of the transformed error term, where the transformation is such that the transformed error is standard normal. We also used products of these cells with a partitioning into four cells of the space of covariates, based upon the value of X_{2i} (total expenditures) only. This yields a partition of the sample space in 20 cells.

In case of ML estimation, the test statistics can be obtained as the explained sum of squares of a regression of a vector $(1, \ldots, 1)' \in \mathbb{R}^N$ on the vectors of scores and the vectors of differences between predicted and sample probabilities for each of the cells (Andrews, 1989, pp. 154–156). Under the null of no misspecification, the test statistics follow a χ^2-distribution, with 16 or 4 degrees of freedom, for 20 and 5 cells, respectively.

The tests were performed for the four parametric specifications in Section 2. In all cases, the hypothesis of no misspecification was strongly rejected. Values of the test statistic for the 20 cells case varied from 84·1 (the normal exponential heteroscedasticity model) to 181·6 (the Tobit model).

ACKNOWLEDGEMENTS

We are grateful to two anonymous referees, Joel Horowitz, Arie Kapteyn, and Theo Nijman for helpful comments, and to the Netherlands Central Bureau of Statistics (CBS) (now named Statistics Netherlands), for providing the data. The views expressed in this paper do not necessarily reflect the policies of the CBS. The first author is grateful to the Netherlands Organization of Scientific Research (NWO) for financial support. The second author's research is made possible by a fellowship of the Netherlands Royal Academy of Arts and Sciences.

REFERENCES

Amemiya, T. (1973), 'Regression analysis when the dependent variable is truncated normal', *Econometrica*, **41**, 997–1016.

Amemiya, T. (1984), 'Tobit models: A survey', *Journal of Econometrics*, **24**, 3–61.
Andrews, D. (1989), 'Chi-square diagnostic tests for econometric models', *Journal of Econometrics*, **37**, 135–156.
Blundell, R. (1990), 'Consumer behaviour: Theory and empirical evidence—A survey', in A. J. Oswald (ed.), *Surveys in Economics*, Cambridge University Press, New York. 1–50.
Blundell, R. (ed.) (1987), 'Specification testing in limited and discrete dependent variable models', *Journal of Econometrics*, **34**, Annals.
Chesher, A., and M. Irish (1987), 'Residual analysis in the grouped and censored normal linear model', *Journal of Econometrics*, **34**, 33–61.
Deaton, A. and J. Muellbauer (1980), *Economics and Consumer Behavior*, Cambridge University Press, New York.
Duncan, G. M. (1986), 'A semi-parametric censored regression estimator', *Journal of Econometrics*, **32**, 5–34.
Eymann, A., and G. Ronning (1992), 'Microeconometric models of tourists' destination choice', mimeo, University of Konstanz.
Fernandez, L. (1986), 'Non-parametric maximum likelihood estimation of censored regression models', *Journal of Econometrics*, **32**, 35–57.
Gabler, S., F. Laisney, and M. Lechner (1993), 'Semi-nonparametric maximum likelihood estimation of binary choice models with an application to labour force participation', *Journal of Business and Economic Statistics*, **11**, 61–80.
Goldberger, S. (1983), 'Abnormal selection bias', in S. Karlin, T. Amemiya, and L. Goodman (eds), *Studies in Econometrics, Time Series, and Multivariate Statistics*, Academic Press, New York.
Horowitz, J. L. (1986), 'A distribution-free least squares estimator for censored linear regression models', *Journal of Econometrics*, **32**, 59–84.
Horowitz, J. L. (1993), 'Semiparametric estimation of a work-trip mode choice model', *Journal of Econometrics*, **58**, 49–70.
Hurd, M. (1979), 'Estimation in truncated samples when there is heteroscedasticity', *Journal of Econometrics*, **11**, 247–258.
Klein, R. L., and R. H. Spady (1993), 'An efficient semiparametric estimator for binary response models', *Econometrica*, **61**, 387–461.
Nelder, J. A., and R. Mead (1965), 'A simplex method for function minimization', *Computer Journal*, **7**, 308–313.
Newey, W. K. (1988), 'Two step series estimation of sample selection models', mimeo, MIT (revised 1991).
Newey, W. K., and J. L. Powell (1990), 'Efficient estimation of linear and type I censored regression models under conditional quantile restrictions', *Econometric Theory*, **6**, 295–317.
OECD (1989), *National and International Tourism Statistics 1974–1985*, Paris.
O'Neill, R. (1971), 'Algorithm AS 47: Function minimization using a Simplex procedure', *Applied Statistics*, **20**, 337–345.
Powell, J. L. (1984), 'Least absolute deviations estimation for the censored regression model', *Journal of Econometrics*, **25**, 303–325.
Powell, J. L. (1986a), 'Censored regression quantiles', *Journal of Econometrics*, **32**, 143–155.
Powell, J. L. (1986b), 'Symmetrically trimmed least squares estimation for Tobit models', *Econometrica*, **54**, 1435–1460.
Powell, J. L. (1987), 'Semiparametric estimation of bivariate latent variable models', mimeo, University of Wisconsin(revised 1989).
Robinson, P. M. (1987), 'Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form', *Econometrica*, **55**, 875–891.
Robinson, P. M. (1988), 'Semiparametric econometrics: A survey', *Journal of Applied Econometrics*, **3**, 35–51.
Ruud, P. A. (1984), 'Tests of specification in econometrics', *Econometric Reviews*, **3**, 211–242.
Ruud, P. A. (1986), 'Consistent estimation of limited dependent variables models despite misspecification of distribution', *Journal of Econometrics*, **32**, 157–187.
SEO (1991), 'A simulation model of the Dutch tourist market', *SEO-report 225*, University of Amsterdam.
Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica*, **26**, 24–36.
Van Soest, A., and P. Kooreman (1987), 'A micro-econometric analysis of vacation behaviour', *Journal of Applied Econometrics*, **2**, 215–226.